ChinaXiv合作期刊

基于交易特征对以太网多类型非法账户的分析与预测 *

周 健 [1,2]，闫 石 [1†]，张 杰 [1]，黄世华 [1]

(1. 安徽财经大学 管理科学与工程学院, 安徽 蚌埠 233040; 2. 北京邮电大学 计算机学院, 北京 100876)

摘 要： 日益频繁的非法交易行为妨害以太坊安全交易，针对电子货币的匿名性使得非法交易行为难于跟踪分析问题。以以太坊平台交易数据作为数据源，以被标记得非法账户和未标记的正常账户数据集作为训练集，利用交易数据的特征属性为构造基础，通过 CatBoost 算法对其中包含多种类型的非法账户进行整体预测。其过程通过 T-SNE 算法实现交易特征的降维可视化，采用多倍交叉验证，引入 SHAP Value 因子判断特征影响的正负属性，所建立模型的预测效果准确率达到了 94.29%，感受者曲线下面积(AUC)数值的评估度量达到了 0.9846。建议的方案较为准确的预测以太坊交易平台上存在的非法行为，有效改善基于区块链的交易环境。

关键词： 区块链；机器学习；以太坊；非法账户；交易特征

中图分类号： TP311.13 doi: 10.19734/j.issn.1001-3695.2022.03.0113

Analysis and prediction of multi-type illegal accounts of ether based on transaction characteristics

Zhou Jian[1, 2], Yan Shi[1†], Zhang Jie[1], Huang Shihua[1]

(1. *College of Management Science & Engineering*, *Anhui University of Finance & Economics*, *Bengbu Anhui* 233040, *China*; 2. *College of Computer Science*, *Beijing University of Posts & Telecommunications*, *Beijing* 100876, *China*)

Abstract: The increasingly frequent illegal transactions hinder the secure transactions of Ethereum, and the anonymity of electronic currency makes it difficult to track and analyze illegal transactions. This paper used the transaction data of the Ethereum platform as the data source, the marked illegal account and unmarked normal account data set as the training set, and the characteristic attributes of the transaction data as the construction basis. Account for the overall forecast. The process uses the T-SNE algorithm to realize the dimensionality reduction and visualization of transaction features, adopts multiple cross-validation, and introduces the SHAP Value factor to judge the positive and negative attributes of the feature. The prediction effect accuracy rate of the established model reaches 94.29%. The evaluation metric for the area (AUC) value reached 0.9846. The proposed scheme can more accurately predict the illegal behavior on the Ethereum trading platform, and will effectively improve the blockchain-based trading environment.

Key words: blockchain; machine learning; ethereum; illegal account; transaction features

0 引言

2008 年区块链技术被中本聪[1]提出，随后以区块链为基础的电子虚拟货币[2,3].如比特币(Bitcoin)[4,5]、以太坊(Ethereum)[6,7]、瑞波币(Ripple)[8,9]等被应用于电子交易中。然而区块链技术的匿名性使得非法交易难于跟踪和分析，这也进一步吸引了犯罪分子，引发了更多的非法交易，如暗网交易[10]、违禁物品交易[11]、金融诈骗[12]等等，全网首个利用区块链智能合约技术[13]实施网络犯罪“BigGame”和”MDF 项目”，涉案数字货币 130 余万个、市值约 2600 余万元。基于区块链技术犯罪成为一种新型高科技犯罪，严重妨害了电子虚拟货币交易的安全性和稳定性。

基于区块链的非法交易分析是一个挑战性问题。目前，非法交易行为的检测仍然以链上数据的分析为主要方法。根据研究方法和研究目标分为三类：a)基于机器学习的数据分析方式，Fan 等人[14]提出了一种许可区块链的隐私保护 DML 模型，以解决其安全性能问题，但在部署和实施上仍然存在不足。到后来，比特币和以太坊中都出现了一种典型欺诈活动--智能庞氏骗局[15]以至于通过采用不同机器学习算法对其进行深入研究以达到预测非法账户的目的，然而其在精度以及探究的问题上依然存在不充分的问题。CHUN WEI[16]等人则使用元启发式算法在区块链环境中找出合适的深度学习超参数，以便进一步探索区块链之间的通信问题，但其在参数的额外通信成本和等待同步等挑战中，仍然存在不足；b)基于复杂网络和交易逻辑结构的非法交易行为分析，如 Chen, Weili[17]对以太坊中存在的智能庞氏骗局合约进行分析，以寻找健康的区块链交易环境，其在数量、数据和研究的非法账户类型上还存在不充分的问题。Dan Lin, Jiajing Wu[18]等人则是通过复杂网络的方法进行建模和理解以太坊中发生的交易信息，以挖掘出潜在的价值交易分析，但其所用方法不是整体的归纳方法，无法添加最新的节点表示；c)基于特征值分析的，如早期通过对比特币钱包[19]进行分析，解析比特币用于大规模犯罪所面临的挑战以及对比特币环境[20,21]中存在的欺诈活动，但在特征提取以及监督方法上存在不足问题。Bartoletti 等人[22]更细入的对以太坊中存在的大量庞氏骗局进行了全面的分析调查，以总结其各类观点的影响，但其在应用层面还缺乏不足，以及对于非法账户的类型仍存在更广泛的探究问题。总结以上这些方法，即准确度距离实际仍依旧存在差距，或者方法在性能上仍然存在优化空间，以及在非法账户预测上只着重关注于某一种类型，例如庞氏骗局[23,24]，钓鱼节点[25,26]，非法洗钱[27,28]等，因此在对区块链上存在非法账户行为的探究仍然存在不足。

收稿日期：2022-03-23；修回日期：2022-05-06 基金项目：国家自然科学基金资助项目(61402001)；安徽省高等学校自然基金资助项目(KJ2020A0013，KJ2019A0657，KJ2018A0441)；安徽财经大学重点项目(ACKY1815ZDB，ACKYB19012)；安徽财经大学科研创新基金资助项目(ACYC2020369)

作者简介：周健，男，安徽蚌埠人，教授，硕导，博士(后)，主要研究方向为智能商务与数据挖掘；闫石，男(通信作者)，安徽合肥人，硕士，主要研究方向为智能商务与数据挖掘(467581941@qq.com)；张杰，男，安徽六安人，硕士，主要研究方向为智能商务与数据挖掘；黄世华，女，河南信阳人，硕士，主要研究方向为智能商务与数据挖掘.

chinaXiv:202206.00072v1

ChinaXiv合作期刊

针对以上存在的方法、性能及预测类型不足等问题，本文在原有交易特征基础上进行创新构造，之后采用机器学习中[29,30]的 K-Means 聚类算法先对数据集中属性特征进行聚类分析，再使用 CatBoost[31,32]进行非法账户的预测。根据划分的特征属性爬取全新的交易动态数据，其数据中的非法账户包含多种类型，并呈无规则排列整理，利用 T-SNE 算法可视化出数据集在属性特征中正常与非法账户的分布状况，构建后的模型在不同参数环境下,根据准确率(Accuracy)，感受者曲线下面积(Receiver Operating Characteristic Curve，AUC)数值这两方面得出最优预测效率的模型结构，通过引入 SHAP Value 变量测出属性特征影响模型构建的正负性，并且本文将与其他机器学习算法进行对比，保证所选方法的高度优良性。实验结果证明，该模型显著提高了对含有多种类型的非法账户预测的正确率。

1　CatBoost 算法

CatBoost 算法作为梯度提升树中最新的研究算法，并未应用到对区块链交易平台上非法账户的预测，其具有能够很好地处理类别特征问题并且有效地减少过拟合问题的特点，根据式(1)将分类特征值转换为数值结果。

$$\hat{x}_k^i=\frac{\sum_{j=1}^{p-1}[x_{\sigma,j,k}=x_{\sigma,p,k}]Y_{\sigma,j}+a\bullet p}{\sum_{j=1}^{p-1}[x_{\sigma,j,k}=x_{\sigma,p,k}]+a} \tag{1}$$

其中：P 是添加的先验项，a 是大于 0 的权重系数，j 是代表类别特征值的系数，k 是训练样本的系数，i 是第 i 个为类别特征，$\sum_{j=1}^{p-1}[x_{\sigma,j,k}=x_{\sigma,p,k}]Y_{\sigma j}$ 是类别特征值中等于标签值的次数，$\sum_{j=1}^{p-1}[x_{\sigma,j,k}=x_{\sigma,p,k}]$ 是总体类别特征值个数。

而 CatBoost 算法作为梯度提升树中的一种，采用对称树作为基学习器，通过一组分类器的串行迭代，得出一个强学习器。CatBoost 的第 k 次迭代目标就是求 h_k，即：

$$h_k=\arg_h\min\frac{1}{m}\sum_{k=1}^{m}[-f_k(x_k,y_k)-h(x_k)]^2 \tag{2}$$

其中：$f_k(x,y)=\frac{\sigma L[y,F_{k-1(x)}]}{\sigma F_{k-1(x)}}$ 为梯度估计，其中 $L[y,F_{k-1(x)}]$ 是损失函数，$F_{k-1(x)}$ 已完成的 k－1 步迭代形成的当前的学习器。为了得到梯度的无偏估计，CatBoost 结合本文数据集，具体建模过程如下：

a)对于非法账户数据集 X 中的每一个样本 x_i，CatBoost 会利用 x_i 之外的全部训练样本并得到模型 M_i

b)采用排序提升利用 M_i 计算 x_i 的梯度估计，即计算 $f_k(x,y)$ 的值。

c)利用新模型对样本 x_i 重新评估并形成一个基学习器。

d)进一步对基学习器进行处理，最终形成强学习器。

以上过程在不断的令非法账户数据集中的 $L[y,F_{k-1(x)}]$ 值变小，即减小模型在训练集中的预测误差，最终形成 CatBoost 模型。

2　非法账户检测

2.1　数据预处理

2.1.1 交易特征

该数据集主要来源于两部分，根据以太坊平台公布的被标记的非法账户数据集以及从平台上爬取下来的正常账户数据集。以太坊社区所提供的非法账户数据集的类型主要包括[33]：试图模仿其他合同提供代币的地址、诈骗彩票、假的初始硬币、模仿其他交易用户、智能庞氏合约骗局和钓鱼节点。选取数据集中非法账户均包含上述多种类型，所建模型针对其包含的多种类型非法账户进行结果预测。

在对正常账户选取时，从以太坊第 1209500 块到 12010000 块之间随机选择了 4024 个正常账户，非法账户则是通过以太坊公开的被标记数据集中选取了 4300 个账户。通过对比收集到的正常账户与非法账户的地址，在筛选后，确保两类账户地址不存在重复，因此总计得到了 4024 个正常账户与 4300 个非法账户。本文通过以太坊提供的 API 进行数据爬取，先是将合计数据传递到 Ethereum 的 API 上，以获得账户所从事的相关交易数据。本文观察 Steven Farrugia 等人[33]筛选出的 43 个特征属性后，进行创新构造，将“min_val_sent”与“max_val_sent”两个交易特征属性进行整合创建新的属性特征，即“Sent_Diff_betweent_max_and_min”再利用这 44 个交易特征属性进行模型构造，如图 1 所示。

序号	提取的属性特征	交易特征描述
1	Avg_min_between_sent_tnx	账户发送交易之间的平均时间（以分钟为单位）
2	Avg_min_between_received_tnx	账户接受交易之间的平均时间（以分钟为单位）
3	Time_Diff_between_first_and_last_(Mins)	第一次和最后一次交易的时间差
4	Sent_tnx	发送的正常交易总数
5	Received_Tnx	收到的正常交易总数
6	Number_of_Created_Contracts	创建的合约交易总数
7	Unique_Received_From_Addresses	账户接受交易的唯一地址总数
8	Unique_Sent_To_Addresses	账户发送交易的唯一地址总数
9	min_value_received	收到的以太币的最小值
10	max_value_received	收到的以太币的最大值
11	avg_val_received	曾经受到的以太币的平均值
12	min_val_sent	曾经发送过的以太币的最小值
13	max_val_sent	曾经发送过的以太币的最大值
14	Sent_Diff_betweent_max_and_min	曾经发送过的以太币差值
15	avg_val_sent	曾经发送过的以太币的平均值
16	min_value_sent_to_contract	发送到合约的以太币的最小值
17	max_val_sent_to_contract	发送到合约的以太币的最大值
18	avg_value_sent_to_contract	发送到合约的以太币的平均值
19	total_transactions_	交易总数
20	total_Ether_sent	为账户地址发送的总以太币
21	total_ether_received	账户地址收到的以太币总数
22	total_ether_sent_contracts	发送到合约地址的总以太币
23	total_ether_balance	执行交易后的以太币总余额
24	Total_ERC20_tnxs	ERC20代币转账交易总数
25	ERC20_total_Ether_received	ERC20代币收到的以太币交易总数
26	ERC20_total_ether_sent	以Ether发送的ERC20代币交易总数
27	ERC20_total_Ether_sent_contract	将ERC20代币转移到其他以太币合的总数
28	ERC20_uniq_sent_addr	发送到唯一账户地址的ERC20代币交易数量
29	ERC20_uniq_rec_addr	从唯一地址收到的ERC20代币交易数量
30	ERC20_uniq_rec_contract_addr	从唯一合约地址收到的ERC20代币交易数量
31	ERC20_avg_time_between_sent_tnx	ERC20代币发送交易之间的平均时间（分钟）
32	ERC20_avg_time_between_rec_tnx	ERC20代币收到交易之间的平均时间（分钟）
33	ERC20_avg_time_between_contract_tnx	发送代币交易之间的ERC20代币平均时间
34	ERC20_min_val_rec	从账户的ERC20代币交易中收到的以太币最小值
35	ERC20_max_val_rec	从账户的ERC20代币交易中收到的以太币最大值
36	ERC20_avg_val_rec	从账户的ERC20代币交易中收到的以太币平均值
37	ERC20_min_val_sent	从账户的ERC20代币交易发送的以太币最小值
38	ERC20_max_val_sent	从账户的ERC20代币交易发送的以太币最大值
39	ERC20_avg_val_sent	从ERC20代币交易为账户发送的以太币平均值
40	ERC20_min_val_sent_contract	从账户发送到合约的ERC20代币交易的最小值
41	ERC20_max_val_sent_contract	从账户发送到合约的ERC20代币交易的最大值
42	ERC20_avg_val_sent_contract	从账户发送到合约的ERC20代币交易的平均值
43	ERC20_uniq_sent_token_name	转移的唯一ERC20代币数量
44	ERC20_uniq_rec_token_name	收到的唯一ERC20代币数量

图 1　完整的特征属性集描述

Fig. 1　Complete feature attribute set description

2.1.2 数据清洗

本文所选取数据集均来自于以太坊提供的实时交易数据，因此复杂的数据交易需要检查缺失值，无效值，空值的存在，以确保模型构建后的准确性和适用性。选择利用 Python 对数据集进行预处理，得知数据集中部分属性特征的确存在缺失值和无效值，如表 1 所示。

表 1　存在缺失值与无效值的属性列

Tab. 1　Check the data

名称	是否存在
ERC20_avg_time_between_sent_tnx	TURE
ERC20_avg_time_between_rec_tnx	TURE
ERC20_avg_time_between_contract_tnx	TURE
ERC20_min_val_sent_contract	TURE
ERC20_max_val_sent_contract	TURE
ERC20_avg_val_sent_contract	TURE

chinaXiv:202206.00072v1

ChinaXiv合作期刊

chinaXiv:202206.00072v1

根据上述结果，对存在问题的属性特征进行处理，其原理是利用属性特征中的整体数据进行平均数的计算，得出的数值再填补到属性特征的缺失值处或无效值处。

为确保数据清洗工作的完善性，通过观察数据集具体分布情况，基于对噪声数据的特征分布有以下特点：a)超过 90%的数据显示为“0”值；b)在整个账户集中测量值与真实值存在较大误差或无法得出其测量值；c)其特征属性会影响模型预测的性能负担；如“total_ether_sent_contracts”，“ERC20_avg_val_sent_contract”等多个特征属性。考虑其或为模型性能负担形数据，因此将此类特征属性进行删除，不引入最终的模型构建。

2.2　数据降维及聚类分析

实验数据集特征空间呈现多维度化，利用 T-SNE 算法在对其进行非线性转换，使其在 2D 平面中视觉出数据集账户标签类型。T-SNE 算法是一种通过二维或三维地图给每个数据点一个位置实现高维数据可视化的统计方法，其算法主要有两大优势：a)对于不相似的点，用一个较小距离会产生较大的梯度来让这些点排斥开来；b)排除不会过于大，即避免不相似的点距离太远。具体算法逻辑如下：

```
算法 1　T-SNE 算法运行伪代码
输入：实验数据集，以及引用 Python 携带的三方库。
输出：T-SNE 可视化后的图形
a)导入实验数据集。
b)去除实验数据集的无关列，即“address”,并将”FLAG“标签数据列作为单独的 y 列。
c)利用 StandardScaler 函数标准化目标特征，并设置 Tsne 算法中的 n_components 参数为 2，以满足二维条件。
d)最后 scatterplot 函数可视化实验效果模型。
```

因此图 2 中表现出，a)标签和非标签数据有几个可区分的集群，且这些集群分散在四周，呈点状式分布排列；b)标签和非标签数据类型的集群任然存在大部分的重叠。c)在右下角明显存在小簇“非法”账户标签集群。以上结果强调了使用机器学习的重要性，以区分两个二进制类型。

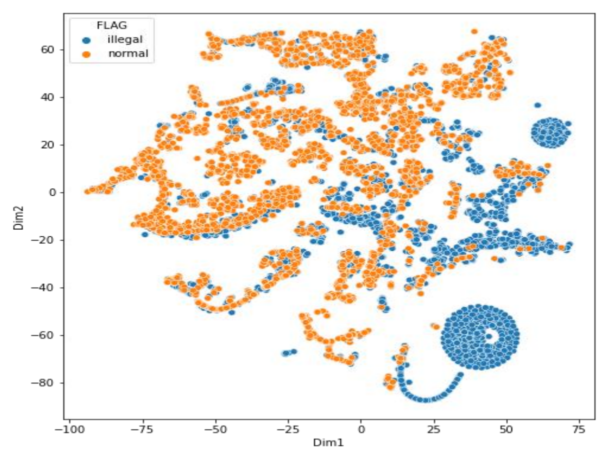


图 2　TSNE 算法显示的标签和非标签数据 2D 分布图

Fig. 2　2D distribution map of labeled and unlabeled data displayed by TSNE algorithm

通过引用 K-Means 聚类算法对数据的类别进行分类归整，其实际意义是将数据集中的属性特征类别进一步聚类分析，观察其不同账户类别中是否处于同一实际效果，为确定 K-Means 算法中 K 值的最优质心，实验选用肘部法则(Elbow Rule)以确定最优 K 的取值，根据肘部法则效果图可以清晰地观察到本文使用的实验数据集最优质心 K 为 2，即当 K=2 时，下降幅度曲线明显趋近于缓慢，具体效果如图 3 所示。

由图 4 的聚类效果可知，实验选取不同账户类别数据集除小部分显示为不同类别特征，大部分均为相同类别特征，进一步强调了实验选用机器学习方法的重要性。

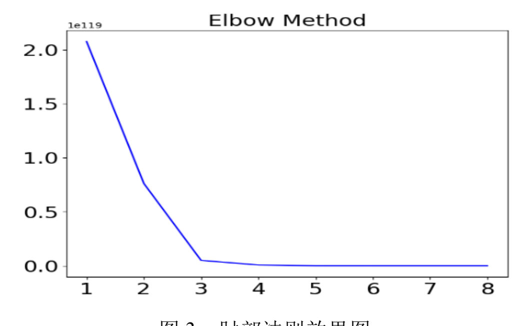


图 3　肘部法则效果图

Fig. 3　Elbow rule renderings

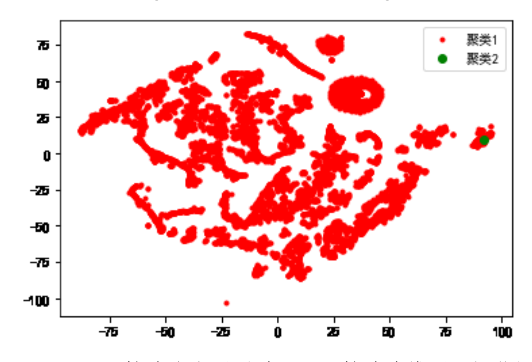


图 4　K-Means 算法聚类后通过 TSNE 算法降维后可视化图形

Fig. 4　The visualization graph given by the dimensionality reduction of the TSNE algorithm after clustering by the kmeans algorithm

3　实验结果分析

3.1　参数评估

通过创建新的交易特征属性以及 K-Means 聚类算法的初步聚类整合，之后建模实验流程均在 python3.9 环境中实现，首先读取实验中的数据集，在使用 Catboost 算法对数据集进行模型建立时，本文考虑了学习率(learning_rate)、学习器的数量(n_estimators)和最大深度(max_depth)三个重要的因素。

由于模型在训练时，无法确定参数最优情况，利用网格搜索优化法对 Catboost 算法进行参数的调整。除上述所选取的三个重要参数因素外，CatBoost 算法的参数还包括最大树数(iterations)，数值型参数的分割数(border_count)，叶的测试方法(leaf_estimation_method)，l2 的正则参数(l2_leaf_reg)等。参数的具体释义如表 2 所示。进行调参时，先保证其他参数保持不变，使用网格搜索法先对 border_count 参数进行调优，进而对 iterations 以及 learning_rate 做进一步的调优，紧接着再确定树的深度(depth)，最终确定模型参数结构。对所选取的参数进行调参后，其优化的具体结果如表 3 所示。

表 2　参数的具体释义

Tab. 2　The specific definition of the parameter

参数	类型	描述
iterations	整数	最大树数
border_count	整数	数值型参数的分割数
l2_leaf_reg	整数	l2 的正则参数
depth	整数	树的深度
learning_rate	浮点型	学习率

根据上述调优结果，利用 Python 环境对 CatBoost 算法在不同交叉倍数验证下得出结果，如表 4 所示，实验结果取同层实验的平均数作为最终数值。

根据表 4 所示，在 3，4，5，11，12 倍交叉验证下最终所得的结果远不如 10 倍交叉验证，10 倍交叉验证有效的加强了模型的构建，以及对以太坊中非法账户预测的精度作出进一步的提升。

ChinaXiv合作期刊

表 3　CatBoost 参数最优情况

Tab. 3　Optimal case of catboost parameters

CatBoost 参数	取值
loss_funcation	logloss
learning_rate	0.1
iterations	800
eval_metric	AUC,Accuracy
border_count	128
l2_leaf_reg	3
rsm	1
depth	5
n_estimators	300
leaf_estimation_method	Gradient
one_hot_max_size	2

表 4　3，4，5，10，11 和 12 倍交叉验证结果

Tab. 4　3, 4, 5,10,11 and 12-fold cross-validation results

交叉验证倍数	树深	学习器数量	Accuracy	AUC
3	7	300	0.9394	0.9824
4	8	250	0.9406	0.9834
5	7	300	0.9410	0.9836
10	5	300	0.9431	0.9851
11	6	300	0.9410	0.9684
12	5	300	0.9401	0.9841

3.2　模型评估效果

利用网格搜索优化法，本文已经确定了三个对构建模型起到至关重要的三个元素：1)学习率；2)树深；3)学习器的数量。而通过上述实验分析，基于 AUC 数值以及对数损失函数(Logloss)对本实验再次细化，根据树深以及学习器数量的不同分别在 10 倍交叉验证下进行模型预测能力评估，其结果以折线图形式表现走向趋势。如图 5 所示。

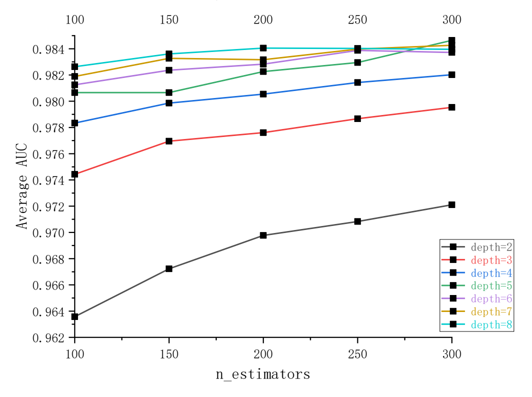


图 5　在 10 倍交叉验证下对 n 个估计量和最大深度参数的验证

Fig. 5　Validation of n estimators and maximum depth parameters under 10-fold cross-validation

通过上述图形可知，树的深度为 2 时，其 AUC 数值在学习器数量增高时明显低于其他深度，但为一个典型的增长趋势，而随着树深值的增大，AUC 的评估度量也在逐步升高，模型参数调整的最优结果也可从图中显示出，当树深在 5(depth=5)，学习器数量抵达 300(n_estimators=300)时,模型此时的 AUC 数值略高于其他参数结果。

而基于参数 Logloss(对数损失)由 CatBoost 算法分别在训练集和测试集中执行迭代次数。从图 6 可观察到，CatBoost 算法在 100 次迭代之后两类数据集同时开始走向收敛，可以证明实验构建的模型，其预测能力具有较好的适应性。

为评估模型的预测能力，从测试集中随机抽取 20 个样本，以观察模型的预测能否达到实际契合效果，如表 5 所示。

表 5 反映出，模型在随机抽取 20 个样本进行预测，所产生的结果与原结果只有一个不同，说明模型具有较强的预测能力。为验证这一结果的可靠性，本文在随机抽取样本的范围上进行扩大，如图 7 所示。

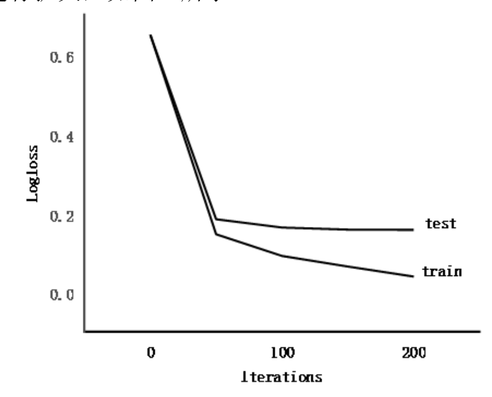


图 6　CatBoost 算法在 10 倍交叉验证下的损失函数

Fig. 6　The loss function of the catboost algorithm under 10 cross-validation

表 5　随机抽取 20 个样本模型预测概率

Tab. 5　Randomly draw 20 samples and predict the probability of the model

抽取样本所在位置	所取样本标签	模型预测所得标签
3064	0	0
1347	1	0
3072	1	1
435	0	0
298	0	0
292	1	1
1729	1	1
4289	1	1
1045	0	0
2427	0	0
2612	1	1
3973	1	1
995	0	0
3919	1	1
211	0	0
379	0	0
3973	0	0
2729	0	0
1018	1	1
534	1	1
总预测概率		95%

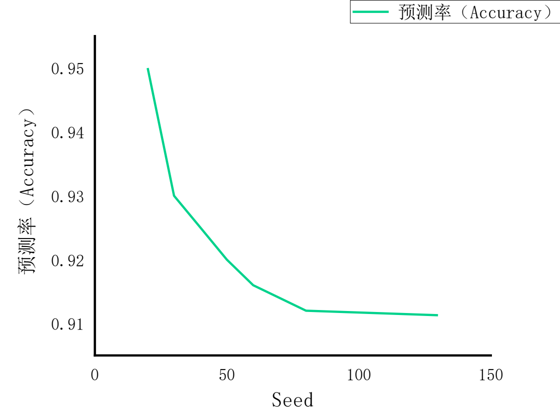


图 7　随机取样范围的预测率变化图

Fig. 7　Prediction rate change graph of the random sampling range

上述图反映出模型预测概率虽然开始时呈现大幅度下降，但随着随机样本的扩大，模型的预测能力逐渐趋于平稳，根据模型最终对整体数据集的预测，预测准确度达到 0.9407，

chinaXiv:202206.00072v1

ChinaXiv合作期刊

chinaXiv:202206.00072v1

这进一步说明本文的模型结构对以太坊中非法账户的预测具有较强的适应性和准确性。

3.3　特征重要性

在构造对以太坊中非法账户预测的模型时，本文也针对所选取数据集中的属性特征进行重要性的排序，因此列举了对模型构建具有较强影响性的十大重要特征，如图 8 所示。

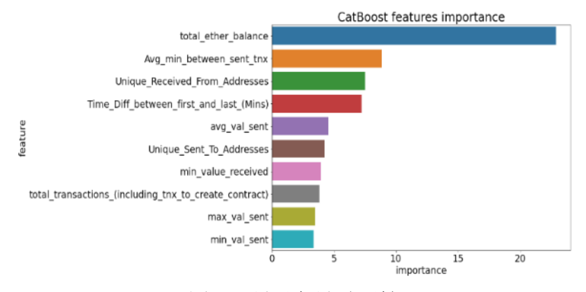


图 8　预测变量重要性

Fig. 8　Predictor importance

通过上述图形结果可知，total_ether_balance,sent_tnx 以及 Unique_Received_from_Address"这个属性特征为显著重要变量，而 Time_Diff_between_first_and_last_(Mins)，Avg_min_between_sent_tnx，min_value_received，Unique_Sent_To_Addresses，avg_val_sent，max_val_sent 和 min_val_sent 这七个预测变量则为一般重要变量，而剩余的 33 个预测变量其重要性所占比值都接近于 0 甚至等于 0，具体的完整预测变量重要性如图 9 所示。

序号	属性名称	重要性排名	影响程度
1	Avg_min_between_sent_tnx	6	4.36%
2	Avg_min_between_received_tnx	9	2.90%
3	Time_Diff_between_first_and_last_(Mins)	4	6.32%
4	Sent_tnx	2	8.14%
5	Received_Tnx	28	0.91%
6	Number_of_Created_Contracts	24	1.10%
7	Unique_Received_From_Addresses	3	7.79%
8	Unique_Sent_To_Addresses	5	5.46%
9	min_value_received	8	4.23%
10	max_value_received	23	1.12%
11	avg_val_received	13	2.34%
12	min_val_sent	11	2.50%
13	max_val_sent	10	2.75%
14	Sent_Diff_betweent_max_and_min	14	2.16%
15	avg_val_sent	7	4.25%
16	min_value_sent_to_contract	44	0.00%
17	max_val_sent_to_contract	43	0.00%
18	avg_value_sent_to_contract	42	0.00%
19	total_transactions_	17	1.73%
20	total_Ether_sent	21	1.29%
21	total_ether_received	12	2.36%
22	total_ether_sent_contracts	34	0.01%
23	total_ether_balance	1	23.47%
24	Total_ERC20_tnxs	16	1.73%
25	ERC20_total_Ether_received	18	1.63%
26	ERC20_total_ether_sent	27	1.02%
27	ERC20_total_Ether_sent_contract	35	0.00%
28	ERC20_uniq_sent_addr	15	1.92%
29	ERC20_uniq_rec_addr	20	1.37%
30	ERC20_uniq_rec_contract_addr	32	0.40%
31	ERC20_avg_time_between_sent_tnx	36	0.00%
32	ERC20_avg_time_between_rec_tnx	37	0.00%
33	ERC20_avg_time_between_contract_tnx	40	0.00%
34	ERC20_min_val_rec	26	1.08%
35	ERC20_max_val_rec	19	1.53%
36	ERC20_avg_val_rec	25	1.09%
37	ERC20_min_val_sent	30	0.47%
38	ERC20_max_val_sent	31	0.44%
39	ERC20_avg_val_sent	29	0.68%
40	ERC20_min_val_sent_contract	38	0.00%
41	ERC20_max_val_sent_contract	39	0.00%
42	ERC20_avg_val_sent_contract	41	0.00%
43	ERC20_uniq_sent_token_name	22	1.13%
44	ERC20_uniq_rec_token_name	33	0.32%

图 9　完整预测变量重要性排名

Fig. 9　Full predictor importance ranking

因此可知在 CatBoost 算法对数据集建模过程中，不同的属性特征对其构造模型的影响是不同的，在对特征属性收集划分中需要整理更多潜在可能影响到模型建造的特征属性，才能使模型的适用性和准确性得到提升。

实验为进一步探究特征属性对样本的影响状况，引入 SHAP Valve 变量，表现出其影响的正负性，具体结果如图 10 所示。

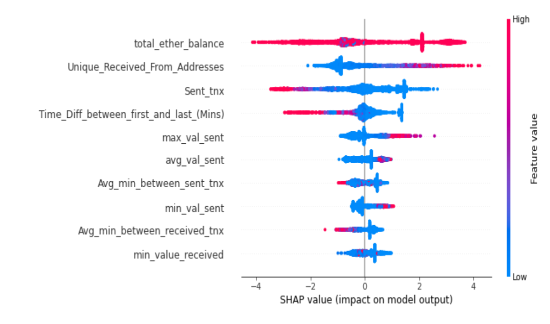


图 10　SHAP value 可视化图

Fig. 10　SHAP value visualization graph

通过上述图形的表现，结合 Feauture Important(属性特征重要性)的结果，可视化出排名前十的特征属性，其中颜色代表特征取值的大小，宽度代表特征分布。而若 SHAP Value 小于零，则对标签特征产生负影响；否则对标签特征产生正影响。图中，"Unique_Received_from_Address"特征属性的正影响因子显著高于负影响因子，然而"sent_tnx"、"Time_Diff_between_first_and_last_(Mins)"这两个特征属性的负影响因子高于正影响因子。通过引入 SHAP Value 变量可直观的判断出不同属性特征对模型构建影响的正负性。

3.4　模型对比

通过交易特征的创建以及其他算法与 CatBoost 的结合，再与其他算法模型和其本身原有模型进行比较，分别使其在 10 倍交叉验证下对实验数据集进行预测，以此验证本文所选用方法的效果较好。具体情况如表 6 所示。

表 6　模型对比实验结果

Tab. 6　Model comparison experiment results

算法模型	精确度(Accuracy)	AUC 数值
决策树	0.906	0.945
SVM(线性)	0.5196	0.764
神经网络	0.7342	0.836
LightGBM	0.7853	0.9695
CAT 树	0.8487	0.888
原 CatBoost	0.9429	0.9846
改进后 CatBoost	0.9431	0.9851

根据表格所示，本文选用的改进后 CatBoost 算法通过交易特征的创建并利用 K-Means 聚类算法先对实验数据集进行聚类整合，再运用 CatBoost 算法进行模型的搭建，其结果无论在精确度或 AUC 数值上均高于其他算法，即本文所构建的模型结构对以太坊中非法账户的预测拥有较好的准确性以及适应性。而列举的其他算法中，与 CatBoost 算法属于同一梯度提升树的 LightGBM 算法，其虽然在 AUC 数值上有良好表现，但其精确度却明显小于 CatBoost 算法。而与单纯的 CatBoost 算法比较，其精确度与 AUC 数值均有略微的提升。因此综上所述，CatBoost 算法通过对比其他算法模型，进一步证明其在对本文收集到的以太坊账户数据集关于非法账户预测上有优异的预测能力，这也使本文所构建的算法模型更具有说服力。

4　结束语

基于对以太坊中公开的账户数据集的搜集与整理，本文利用 Python 开发环境使用 CatBoost 算法对整理的数据集进行模型构建，所构建模型具有预测未来以太坊中非法账户的能力。构建模型不仅在精确度以及 AUC 数值方面有较高水准，还展现了其预测变量重要性的顺序以及具体对模型有正负影响力的能力。本文所提出的算法模型具有较强的适应性，方法也建立在基础层面，对其他应用领域均有预测作用，在本文的基础上可以通过进一步的优化和改进以探究区块链交

ChinaXiv合作期刊

易平台中更深层次的非法行为。根据以太坊中海量账户之间的交易历史来检测是否存在非法账户，从而为这一研究领域作出了贡献，所提出的算法模型也可以被相关的经济机构和部门所使用。

关于未来的工作，团队会有的架构模型在其他应用领域进行预测，以保证模型的适用性和可实践性，对于不仅存在在以太坊中的非法账户预测，而且其他关于基于区块链的交易平台非法预测，团队都将进行分析与探究。在关于属性特征上，会进一步尝试提取区块链交易中的新型特征，以完善模型在预测非法账户上的能力。本文使用 CatBoost 算法对模型进行搭建，其本身也存在或多或少的缺陷，会通过优化算法对模型预测能力的准确度上进行提升，使模型更具效率和说服力。

参考文献：

[1] Nakamoto S. Bitcoin: A peer-to-peer electronic cash system. White Paper. 2008

[2] Tschorsch F, Scheuermann B. Bitcoin and Beyond: A Technical Survey on Decentralized Digital Currencies [J] , IEEE Communications Surveys and Tutorials. 2016, 18 (3): 2084-2123.

[3] Chipere M. Virtual currency as an inclusive monetary innovation for the unbanked poor [J] , Electronic Commerce Research and Applications. 2018, 28: 28-37

[4] Bohme R, Christin, N, Edelman, B, *et al.* Bitcoin: Economics, Technology, and Governance [J] , Journal of Economic Perspective. 2015, 29 (2): 213-238

[5] Sajana P, Sindhu M, Sethumadhavan M. On blockchain applications: Hyperledger fabric and Ethereum [J] , International Journal of Pure and Applied Mathematics. 2018, 118 (18): 2965-2970.

[6] Wang S, Ouyang, LW, Yuan Y, *et al.* Blockchain-Enabled Smart Contracts: Architecture, Applications, and Future Trends [J] , IEEE Trans on Systems Mam Cybernetics-Systems. 2019, 49 (11): 2266-2277

[7] Hasan M, Naeem MA, Arif M, *et al.* Higher moment connectedness in cryptocurrency market [J] , Journal of Behavioral and Experimental Finance. 2021, 32.

[8] Kristoufek L, Vosvrda M, *et al.* Cryptocurrencies market efficiency ranking: Not so straightforward [J] , Physica-Statistical Mechanics and Its Applications. 2019, 531.

[9] ElBahrawy A, Alessandretti L, Rusnac L, *et al.* Collective dynamics of dark web marketplaces [J] , Scientific Reports. 2020, 10 (1) .

[10] Lehdonvirta V. Virtual item sales as a revenue model: identifying attributes that drive purchase decisions [J] , Electronic Commerce Research. 2009, 9 (1-2): 97-113.

[11] Hyvarinen H, Risius M, Friis G, *et al.* A Blockchain-Based Approach Towards Overcoming Financial Fraud in Public Sector Services [J] , Business & Information Systems Engineering. 2017, 59 (6): 441-456.

[12] Bayramova A, Edwards DJ, Roberts C, *et al.* The Role of Blockchain Technology in Augmenting Supply Chain Resilience to Cybercrime [J] , Buildings, 2021, 11 (7) .

[13] Kim H, Kim SH, Hwang JY, *et al.* Efficient Privacy-Preserving Machine Learning for Blockchain Network [J] , IEEE ACCESS. 2019.

[14] Fan SH, Fu SJ, Xu HR, *et al.* Anti-leakage smart Ponzi schemes detection in blockchain [J] , Information Processing & Management. 2021, 58 (4) .

[15] Tsai CW, Chen YP, Tang TC, *et al.* An efficient parallel machine learning-based blockchain framework [J] , ICT Express. 2021, 7 (3) , 300-307.

[16] Chen, Weili. Detecting. ponzi schemes on ethereum: Towards healthier blockchain technology [C]// Proceedings of the 2018 World Wide Web Conference. 2018.

[17] Dan Lin, Jiajing Wu, Qi Yuan, *et al.* Modeling and Understanding Ethereum Transaction Records via a Com-plex Network ApproachP [J] , IEEE Trans on Circuits And Systems Novemeber. 2020.

[18] Meiklejohn S, Pomarole M, Jordan, G, *et al.* A fistful of bitcoins: characterizing payments among men with no names [C]// Proceedings of The 2013 Conference On Internet Measurement Conference. 2013.

[19] Monamo, P, Vukosi M, *et al.* Unsupervised learning for robust Bitcoin fraud detection [C]// 2016 Information Security for South Africa (ISSA) . IEEE, 2016.

[20] Abad-Segura E, Infante-Moro A, Gonzalez-Zamar MD, *et al.* Blockchain Technology for SecureAccounting Management: Research Trends Analysis [J] , Mathematics, 2021, 9 (14) .

[21] Bartoletti, M, Carta S, Cimoli T, *et al.* Dissecting Ponzi schemes on Ethereum: identification, analysis, and impact [J] , Future Generation Computer Systems. 2020, 101: 259-277.

[22] Bartoletti M, Carta S, Cimoli T, *et al.* Dissecting Ponzi schemes on Ethereum: Identification, analysis, and impact [J] , Future Generation Computer Systems-The International Journal Of Escience. 2020, 102: 257-277.

[23] Gurun UG, Stoffman N, Yonker SE. Trust Busting: The Effect of Fraud on Investor Behavior [J] , Review of Financial Studies, 2018, 31 (4): 1341-1376.

[24] Chen L, Peng JY, Liu Y, *et al.* Phishing Scams Detection in Ethereum Transaction Network [J] , ACM Trans on Internet Technology. 2021, 21 (1) .

[25] Lin GY, Liu BW, Xiao PC, *et al.* Phishing Detection with Image Retrieval Based on Improved Texton Correlation Descriptor [J] , Cmc-Computers Materials & Continue. 2018, 57 (3): 533-547.

[26] Campbell-Verduyn M. Bitcoin, crypto-coins, and global anti-money laundering governance [J] , Crime Law And Social Change. 2018, 69 (2) , 283-305.

[27] Ducas E, Wilner A. The security and financial implications of blockchain technologies: Regulating emerging technologies in Canada [J] , International Journal. 2017, 72 (4): 538-562.

[28] Bradley, AP. The use of the area under the roc curve in the evaluation of machine learning algorithms [J] , Pattern Recognition. 1997.

[29] 周建，张杰，闫石. 基于链上数据的区块链欺诈账户检测研究 [J]. 计算机应用研究, 2022, 39 (04): 992-997. (Zhou Jian, Zhang Jie, Yan Shi. Research on detection of fraudulent accounts in blockchain based on on-chain data [J] , Application Research of Computers. 2022, 39 (04): 992-997.)

[30] 俞莎莎，牛保宁. 基于交易不可信度的比特币非法交易检测 [J]. 计算机工程, 2021. (Yu Shasha, Niu Baoning. Bitcoin illegal transaction detection based on transaction unreliability [J]. Computer Engineering, 2021.

[31] Huang GM, Wu LF, Ma X, *et al.* Evaluation of CatBoost method for predic tion of reference evapotranspiration in humid regions [J] , Journal of Hydrology. 2019. 574: 1029-1041.

[32] Punmiya R, Choe S. Energy Theft Detection Using Gradient Boosting Theft Detector With Feature Engineering-Based Preprocessing [J] , IEEE Trans on Smart Grid, 2019, 10 (2): 2326-2329.

[33] Steven F, Joshua E, George A. Detection of illicit accounts over the Ethereum blockchain[J], Expert Systems with Applications.2020,15.

chinaXiv:202206.00072v1